Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Spline-based Self-controlled Case Series Method

Yonas Ghebremichael-Weldeselassie [a], Heather J. Whitaker[a*†], and C. Paddy Farrington[a]

The self-controlled case series (SCCS) method is an alternative to study designs such as cohort and case control methods and is used to investigate potential associations between the timing of vaccine or other drug exposures and adverse events. It requires information only on cases, individuals who have experienced the adverse event at least once, and automatically controls all fixed confounding variables that could modify the true association between exposure and adverse event. Time-varying confounders such as age, on the other hand, are not automatically controlled and must be allowed for explicitly. The original SCCS method used step functions to represent risk periods (windows of exposed time) and age effects. Hence exposure risk periods and/or age groups have to be pre-specified a priori, but a poor choice of group boundaries may lead to biased estimates. In this paper, we propose a non-parametric SCCS method in which both age and exposure effects are represented by spline functions at the same time. To avoid a numerical integration of the product of these two spline functions in the likelihood function of the SCCS method we defined the first, second and third integrals of I-splines based on the definition of integrals of M-splines. Simulation studies showed that the new method performs well. This new method is applied to data on paediatric vaccines.

Keywords: Integral of I-splines; M-splines; Non-parametric SCCS; Smooth risk functions.

1. Introduction

The self-controlled case series (SCCS) method is used to investigate potential associations between transient exposures and adverse health events [1]. It yields estimates of relative incidence, that is, the incidence in exposure risk periods relative to all other time over which cases are observed. The method uses information only from cases, individuals who have experienced the event of interest at least once, and implicitly controls all measured and unmeasured confounding variables that act multiplicatively on the hazard. However, time varying confounders such as age are not automatically controlled and hence should be included in the model. Detailed description of the self-controlled case series method can be found in [1], [2] and [3]. The standard SCCS method uses piecewise constant step functions to represent both age and exposure effects. Poor choice of the a priori chosen age groups or exposure risk periods in the standard method may result in biased estimates of exposure-related relative incidences. Usually the choice of exposure risk periods is motivated by reference to previous studies or hypotheses, by biologically plausible mechanisms or by expert opinion, but it is not

[a] *Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK*
[*] *Correspondence to: Heather Whitaker: Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.*
[†] E-mail: heather.Whitaker@open.ac.uk

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

uncommon to face a situation in which there is little knowledge of the precise timing defining true exposure risk periods. Recently the SCCS method was extended by modeling the age effect using splines to avoid the limitations of the standard SCCS model [4]. In addition linear combinations of cubic M-splines (piecewise polynomials of degree three) were used to represent the exposure risk effect [5]. However, in these two extensions either age or exposure risk effects are represented by piecewise constant functions requiring a priori choice of cut points. Therefore, in this paper we extend the SCCS method further by modeling both age and exposure effects using splines to have a fully non-parametric SCCS method. However, this extension is non-trivial methodologically as it involves integrals of spline products that had not previously been evaluated. We thus provide a new, fully nonparametric SCCS method of analysis.

The new method is applied to data on measles, mumps and rubella (MMR) vaccine and febrile convulsions. The new results are compared with results previously obtained in [1] and additional standard SCCS analyses with varying exposure risk periods.

The paper is organized as follows; after some initial remarks in Section 2, the likelihood function of the spline-based SCCS method is derived in Section 3. In this section, we also describe and define derivatives and integrals of M and I splines, and the integral of a product of two spline functions. Section 4 presents the penalized log-likelihood function of the spline-based SCCS method and discusses the selection of smoothing parameters. In Section 6, we evaluate the performance of the new method using simulations. We apply the spline-based SCCS method to data on febrile convulsions and MMR vaccine in Section 7 and follow this with final remarks in Section 8.

2. Modelling Age and Exposure Effects Using Splines

The use of regression splines in the context of the self-controlled case series method has shown an improved performance compared to the use of step functions [4], [5]. Among the motivations for using regression splines based on M-splines in these papers were that the spline functions give flexible and plausible shapes of age and exposure-related relative incidence functions and avoid numerical integration of the integral in the denominator of the SCCS likelihood function. This numerical integration is avoided because the integral of an M-spline is an I-spline, therefore the integral of a linear combination of M-splines can be expressed as a linear combination of I-splines. Based on similar arguments, both age and exposure effects can be represented as linear combinations of M-spline basis functions. In this paper, since age and exposure are to be represented by linear combinations of M-splines at the same time, the denominator of the SCCS likelihood function will involve the integral of a product of two spline functions. This cannot be represented by a linear combination of I-splines only, so the integration cannot be avoided in the same way as in [4] and [5]. Therefore, based on the definition of the integral of an M-spline developed by Ramsay [6], we define first, second and third integrals of an I-spline to avoid numerical integration of the product of two spline functions. In the following section we derive the likelihood function of the SCCS method when both age and exposure effects are approximated by linear combinations of M-spline basis functions.

3. Likelihood Function

3.1. The general SCCS likelihood

We begin with the general SCCS likelihood function [7]. The likelihood is specified over an observation period defined by age or calendar time boundaries $(a_i, b_i]$ within which an event has been observed and the full exposure history is known. Note that in this paper we take the underlying time line as age, while in practice this can be replaced with calendar time.

The likelihood function of the SCCS method may be derived from a cohort model based on the following three assumptions: (1) events arise in a non-homogenous Poisson process; (2) the occurrence of an event must not alter the

probability of subsequent exposure; and (3) censoring of individuals at the end of the observation period occurs completely at random, so that the occurrence of the event of interest must not censor or affect the observation period [1, 7]. Departures from these assumptions are discussed in [8, 9].

Suppose that individual i, $i = 1, 2, \ldots, N$, from a cohort is observed over age intervals $(a_i, b_i]$, their observation periods, that may vary between individuals in the cohort and $(a_i, b_i] \subseteq (a, b]$. Let x_i^t be the history of exposure and observation up to age t. Therefore $x_i = x_i^{b_i}$ is the exposure and observation history up to the end of observation period. Assume that events of interest for individual i arise in a non-homogeneous Poisson process with intensity function $\lambda(t|x_i^t)$. Therefore the cohort likelihood contribution L_{ci} of an individual case i who has experienced $n_i > 0$ events at ages $t_{i1}, t_{i2} \ldots t_{in_i}$ in the observation period $(a_i, b_i]$ is given by:

$$L_{ci} = \prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i^{t_{ij}}) \exp\left\{ - \int_{a_i}^{b_i} \lambda_i(s|x_i^s) ds \right\}$$

The SCCS likelihood L, up to a multiplying constant, is then obtained from this cohort likelihood function by conditioning on the total number of events, n_i, experienced by an individual i and his/her exposure and observation history during their observation period $(a_i, b_i]$.

$$L = \prod_{i=1}^{N} \frac{\prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i)}{\left\{ \int_{a_i}^{b_i} \lambda_i(s|x_i) ds \right\}^{n_i}}$$

A convenient way of parameterizing the incidence $\lambda(t|x_i)$ is according to the proportional incidence model [7]

$$\begin{aligned} \lambda_i(t|x_i) &= \lambda_0(t) \exp\{\gamma_i + x_i(t)\beta\} \\ &= \varphi\psi(t) \exp\{\gamma_i + x_i(t)\beta\}, \end{aligned}$$

where $\lambda_0(t) = \varphi\psi(t)$ is the baseline incidence at age t, $\psi(t)$ is the age related relative incidence function, φ is the underlying incidence at age a, γ_i is a sum of fixed and random individual effects, $x_i(t)$ is the exposure status at age t, a binary variable when there is only one exposure risk period. The main parameter of interest is $\exp(\beta)$, the exposure relative incidence, that is, the incidence during exposure risk periods relative to all other periods observed before and after exposure.

Therefore parametrization of the incidence function this way gives an SCCS likelihood function for one exposure risk period as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} \psi(t) \exp\{x_i(t)\beta\} dt} \tag{1}$$

Equation 1 can be generalized as follows by using time since start of exposure as an argument for the exposure effect:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij})\omega(t_{ij} - c_i)}{\int_{a_i}^{b_i} \psi(t)\omega(t - c_i) dt}, \tag{2}$$

where $\omega(t - c)$ is the exposure-related relative incidence function which takes the value one if the event falls before the start of exposure (c_i) or after the end of exposure (d_i). In the standard SCCS method, $\psi(t)$ and $\omega(t - c)$ are represented by step functions; in the semi-parametric version of SCCS [7], $\psi(t)$ is left unspecified and $\omega(t - c)$ is fitted as a step function; in [4], $\psi(t)$ is approximated by spline functions and $\omega(t - c)$ by a step function, and in [5], $\psi(t)$ is represented as a step function and $\omega(t - c)$ as a linear combination of M-spline functions.

The identifiability of age and exposure effects has been discussed in detail in the case of the semiparametric SCCS model [7], and these conditions are also required here. The key requirement is that there must be variation in exposure between cases. If there is no such variation, then $c_i = c$ for all i, and $\psi(t)$ and $\omega(t - c)$ cannot be separately identified. This

arises only in very special circumstances, for example those involving shared environmental exposures: in such situations, short observation periods are used in which it is reasonable to assume that there are no substantial age effects. When there is little variation in exposure timing between exposed cases, or when the exposure periods are very long, the inclusion of unexposed cases helps to reduce any confounding between the effects of age and exposure.

3.2. Spline-based SCCS likelihood function

In this paper, we approximate both $\psi(t)$ and $\omega(t-c)$ as linear combinations of cubic M-spline basis functions. M-splines, which are variants of B-splines, are piecewise polynomials connected at points known as knots. Over the domain in the interval $[a,b]$, an M-spline of order q can be defined using the knot sequence $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$, where $k_q = a$ and $k_{q+s+1} = b$ and s is the number of interior knots. Therefore l^{th} M-spline of order q is defined as

$$M_l(t|q) = \begin{cases} \frac{q\{(t-k_l)M_l(t|q-1)+(k_{l+q}-t)M_{l+1}(t|q-1)\}}{(q-1)(k_{l+q}-k_l)}, & k_l \le t < k_{l+q} \\ 0, & \text{elsewhere,} \end{cases}$$

with

$$M_l(t|1) = \begin{cases} \frac{1}{(k_{l+1}-k_l)}, & k_l \le t < k_{l+1} \\ 0, & \text{elsewhere.} \end{cases}$$

where $l = 1, 2, \ldots m$ and $m = q + s$ is the number of M-spline basis functions for a given sequence of knots.

The integrals of M-splines have been defined as I-splines [6]. I-splines are piecewise polynomials of order $q+1$ obtained by integrating M-splines of order q and are thus defined for $k_h \le t < k_{h+1}$ as $I_l(t|q) = \int_o^t M_l(u|q)du$, where the lower limit of the integral is the minimum interior knot denoted by o and $h = 1, 2, \ldots, 2q+s$.

Thus for the same sequence of interior knots used in defining M-splines, I-splines are defined as

$$I_l(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h}(k_{m+q+1} - k_m)\frac{M_m(t|q+1)}{q+1}, & h-q+1 \le l \le h \\ 1 & l < h-q+1. \end{cases}$$

$\psi(t)$ is defined between $a = \min\{a_i; i = 1, \ldots, N\}$ and $b = \max\{b_i; i = 1, \ldots, N\}$, where N is the total number of cases in the study. Since $\psi(t)$ is a relative effect it has to be a positive function and to obtain such a function based on M-splines we constrain the coefficients to be non-negative to give the following expression for $\psi(t)$:

$$\psi(t) = \sum_{l=1}^{m_1} g(\alpha_l) M_{1l}(t) = \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t). \tag{3}$$

The $g(\alpha_l)$ are parameters used to determine the shape of $\psi(t)$ and are constrained to be non-negative by taking $g(\alpha_l) = \alpha_l^2$. $M_{1l}(t)$ is the l^{th} M-spline basis function related to age, m_1 is the number of parameters or the number of M-spline basis functions which is equal to the sum of the number of interior knots and the order of the basis functions.

Similarly, the exposure-related relative incidence function with non-negative coefficients is defined between 0 and $\max\{(d_i - c_i); i = 1, \ldots, N\}$ (in terms of time since risk start), where c_i and d_i are ages at the start and end of exposure related risk respectively for individual i. The risk period is a period where the exposure-related relative incidence can be different from 1; outside this period the exposure-related relative incidence function takes the value 1. Therefore, the exposure-related relative incidence function is defined as:

$$\omega(t-c) = \begin{cases} \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c), & c \le t \le d \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

where m_2 is the number of M-spline basis functions used to define the exposure-related relative incidence function $\omega(t-c)$ and $M_{2l}(t-c)$ is the l^{th} basis function related to exposure. The knots which are used to define the M-splines related to the age effect and the exposure effect are chosen to be equidistant including the arbitrary knots added below and above the minimum and maximum values of the variable.

There are a few points to note about the way risk periods are defined. The overall length of the risk period $(c_i, d_i]$ can vary between individuals. When the exposure is a point exposure, e.g. a vaccine, a nominal post-exposure risk period is defined, thus $d_i = c_i$ plus the length of nominal risk period. If d_i falls beyond the end of observation, b_i, $d_i = b_i$. Risk periods can be unbounded to the right thus ending at b_i. When representing the exposure-related risk with a flexible function it is better to choose the risk period to be too long rather than too short, a simulation study showed this to have little impact on relative incidence estimates [5].

Now replacing $\psi(t)$ and $\omega(t-c)$, in Equation (2), by the spline functions in Equations (3) and (4) respectively gives the likelihood function for the spline-based SCCS as

$$L = \prod_{i=1}^{N}\prod_{j=1}^{n_i} \frac{\left\{\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t_{ij})\right\}\left\{\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t_{ij}-c_i)\right\}^{I(c_i\le t_{ij}\le d_i)}}{\int_{a_i}^{b_i}\left\{\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t)\right\}\left\{\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t-c_i)\right\}^{I(c_i\le t\le d_i)}dt} \tag{5}$$

and the log-likelihood function is

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left[\frac{\left\{\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t_{ij})\right\}\left\{\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t_{ij}-c_i)\right\}^{I(c_i\le t_{ij}\le d_i)}}{\int_{a_i}^{b_i}\left\{\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t)\right\}\left\{\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t-c_i)\right\}^{I(c_i\le t\le d_i)}dt}\right]. \tag{6}$$

To further simplify the denominator of the log-likelihood function (6) and to avoid numerical integration, we use integration by parts. This involves derivatives and integrals of linear combinations of M-spline functions and integrals of their integrals. Therefore, before we proceed with simplifying the log-likelihood function, we define integrals of I-splines in the following subsections. Derivatives of M-splines are described in appendix A1.

3.2.1. Integrals of I-splines Based on the definition for the integral of an M-spline [6] we define the integral of an I-spline. Let the integral of $I_l(t|q)$ be denoted by $I_l^1(t|q) = \int_o^t I_l(u|q)du$. Using the same sequence of interior knots employed to define the M-splines, for $k_h \le t < k_{h+1}$ the integral of an I-spline, $I_l^1(t|q)$, has three different expressions depending on the value of l. For $l > h$ the value of an I-spline is zero so its indefinite integral will be a constant, and hence

$$I_l^1(t|q) = \int_o^t I_l(u|q)du = 0.$$

For $h-q+1 \le l \le h$ an I-spline, $I_l(t|q)$, is given by

$$I_l(t|q) = \sum_{m=l}^{h}(k_{m+q+1}-k_m)\frac{M_m(t|q+1)}{q+1}$$

therefore its integral will be

$$\begin{aligned} I_l^1(t|q) &= \int_o^t \sum_{m=l}^{h}(k_{m+q+1}-k_m)\frac{M_m(u|q+1)}{q+1}du \\ &= \sum_{m=l}^{h}\frac{(k_{m+q+1}-k_m)}{q+1}\int_o^t M_m(u|q+1)du. \end{aligned}$$

$\int_o^t M_m(u|q+1)du$ in the above expression is the integral of an M-spline of order $q+1$ that gives another I-spline, $I_m(t|(q+1)) = \sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}$ for $h-q \leq m \leq h$, so

$$I_l^1(t|q) = \sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}.$$

For $l < h - q + 1$, that is for any value of $t > k_{l+q}$ the value of $I_l(t|q) = 1$. This is because $M_l(t|q) = 0$ for all values of $t > k_{l+q}$. Now the integral of $I_l(t|q)$ has two parts for $t > k_{l+q}$, the integral of the function up to k_{l+q} and from k_{l+q} to t. That is,

$$\int_o^{k_{l+q}} I_l(u|q)du + \int_{k_{l+q}}^{t} I_l(u|q)du = \left\{\sum_{m=l}^{h} \frac{(k_{m+q+1} - k_m)}{q+1} \int_o^{k_{l+q}} M_m(u|q+1)du\right\} + (t - k_{l+q}).$$

Therefore, in summary the integral of an I-spline is given by

$$I_l^1(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}, & h-q+1 \leq l \leq h \\ t - k_{l+q} + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(k_{l+q}|q+2)}{q+2}, & l < h-q+1. \end{cases}$$

The second integral of an I-spline $I_l^2(t|q) = \int_o^t I_l^1(u|q)du$ and the third integral $I_l^3(t|q) = \int_o^t I_l^2(u|q)du$ can be obtained in a similar way (see appendix Section A2).

3.2.2. Integrating the Product of Two Spline Functions Now going back to the log-likelihood function (6), since the exposure-related relative incidence function, $\omega(t-c)$, takes the value 1 in the control periods, $(a_i, c_i]$ and $(d_i, b_i]$, within the observation period, the denominator of the log-likelihood function can be rewritten as

$$\int_{a_i}^{c_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt + \int_{c_i}^{d_i} \left\{\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right\}\left\{\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c_i)\right\} dt + \int_{d_i}^{b_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt$$

Furthermore, the first and the last terms are integrals of only one function, the age-specific relative incidence $\psi(t)$, whereas the second term is the integral of a product of two spline functions. Since the integral of an M-spline of order q is an I-spline of order $q+1$, hence the integral of a linear combination of M-splines can be expressed as a linear combination of I-splines. Therefore, we replace the integrals in the first and third terms by linear combinations of I-spline basis functions which leads to a denominator with the expression

$$\sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)|_{a_i}^{c_i} + \int_{c_i}^{d_i} \left\{\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right\}\left\{\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c_i)\right\} dt + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)|_{d_i}^{b_i}.$$

The $I_{1l}(t)$ are I-splines related to the age effect and $I_{2l}(t)$ will be used to denote I-splines related to the exposure effect. The remaining part in the denominator of the log-likelihood function of the spline-based SCCS is in the risk period $(c_i, d_i]$ where the exposure-related relative incidence can take a value different from 1. This part contains an integral of the product of the two spline functions, $\psi(t)$ and $\omega(t-c)$.

To evaluate this integral we use integration by parts as follows:

$$\int \psi(t)\omega(t-c)dt = \psi(t)\int \omega(t-c)dt - \int \left\{\psi'(t)\int \omega(t-c)dt\right\} dt \quad (7)$$

where $\psi'(t)$ is the first derivative of $\psi(t)$. Since $\psi(t)$ and $\omega(t-c)$ are linear combinations of M-spline basis functions, $\int \omega(t-c)dt$ can be expressed as a linear combination of I-splines denoted by $I_E(t-c)$

$$I_E(t-c) = \int_c^t \omega(u-c)du = \int_c^t \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(u-c)du = \sum_{l=1}^{m_2} \beta_l^2 I_{2l}(t-c).$$

Letting the integral of the linear combination of I-splines $I_E(t-c)$ be denoted by $I_E^1(t-c)$, the integral of $I_E^1(t-c)$ by $I_E^2(t-c)$ and the integral of $I_E^2(t-c)$ by $I_E^3(t-c)$ that is,

$$I_E^1(t-c) = \int I_E(t-c)dt, \quad I_E^2(t-c) = \int I_E^1(t-c)dt \quad \text{and} \quad I_E^3(t-c) = \int I_E^2(t-c)dt,$$

the expression in Equation (7) becomes

$$\int \psi(t)\omega(t-c)dt = \psi(t)I_E(t-c) - \int \{\psi'(t)I_E(t-c)\}\, dt.$$

The last term of this equation is again an integral of a product of two non-constant functions. We therefore apply integration by parts repeatedly until none of the terms is an integral of two non-constant functions and get:

$$\begin{aligned} \int \psi(t)\omega(t-c)dt &= \psi(t)I_E(t-c) - \int \{\psi'(t)I_E(t-c)\}\, dt \\ &= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \psi''(t)I_E^2(t-c) - \psi'''(t)I_E^3(t-c) \end{aligned}$$

where $\psi'(t)$, $\psi''(t)$ and $\psi'''(t)$ are the first, second and third derivatives of $\psi(t)$ respectively. $\psi'''(t)$ is a constant function because $\psi(t)$ is a piecewise cubic function.

3.2.3. Full spline-based SCCS likelihood The log-likelihood function of the spline-based SCCS method, obtained by replacing the appropriate expressions for the terms $\int_{a_i}^{c_i} \psi(t)dt$, $\int_{c_i}^{d_i} \psi(t)\omega(t-c)dt$ and $\int_{d_i}^{b_i} \psi(t)dt$ in the denominator, is

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left[\frac{\left\{\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t_{ij})\right\}\left\{\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t_{ij}-c_i)\right\}^{I(c_i<t_{ij}\le d_i)}}{B}\right] \tag{8}$$

where

$$\begin{aligned} B &= \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\Big|_{a_i}^{c_i} + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(t)\Big|_{d_i}^{b_i} \\ &\quad + \left\{\psi(t)I_E(t-c_i) - \psi'(t)I_E^1(t-c_i) + \psi''(t)I_E^2(t-c_i) - \psi'''(t)I_E^3(t-c_i)\right\}\Big|_{c_i}^{d_i} \end{aligned}$$

and

$$\begin{aligned} I_E^1(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^1(t-c), \\ I_E^2(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^2(t-c), \\ I_E^3(t-c) &= \sum_{l=1}^{m_2} \beta_l^2 I_{2l}^3(t-c) \end{aligned}$$

where $I^1_{2l}(t-c)$, $I^3_{2l}(t-c)$ and $I^3_{2l}(t-c)$ are the first, second and third integrals of the l^{th} I-spline $(I_{2l}(t-c))$ related to exposure, respectively.

So far, the methodology developed in this paper has considered only one exposure risk period. However, it can be applied to multiple exposures provided that the exposure risk periods do not overlap. The relative incidence will be represented by the age-specific relative incidence function in the control periods, and by the product of the age-specific and exposure-specific relative incidence functions during the multiple risk periods. For example, if we have a second non-overlapping exposure risk period $(e_i, f_i]$, the numerator of the log-likelihood function (8) will be multiplied by a spline function related to the new exposure, $\left(\sum_{l=1}^{m_3}\gamma_l^2 M_{3l}(t_{ij}-e_i)\right)^{I(e_i<t_{ij}\leq f_i)}$ during this period. The denominator then becomes

$$\begin{aligned} B &= \sum_{l=1}^{m_1}\alpha_l^2 I_{1l}(t)\big|_{a_i}^{c_i} + \sum_{l=1}^{m_1}\alpha_l^2 I_{1l}(t)\big|_{d_i}^{e_i} + \sum_{l=1}^{m_1}\alpha_l^2 I_{1l}(t)\big|_{f_i}^{b_i} \\ &\quad + \left\{\psi(t)I_E(t-c_i) - \psi'(t)I^1_E(t-c_i) + \psi''(t)I^2_E(t-c_i) - \psi'''(t)I^3_E(t-c_i)\right\}\big|_{c_i}^{d_i} \\ &\quad + \left\{\psi(t)I_{E2}(t-e_i) - \psi'(t)I^1_{E2}(t-e_i) + \psi''(t)I^2_{E2}(t-e_i) - \psi'''(t)I^3_{E2}(t-e_i)\right\}\big|_{e_i}^{f_i} \end{aligned}$$

where $I_{E2}(t-e_i)$, $I^1_{E2}(t-e_i)$, $I^2_{E2}(t-e_i)$ and $I^3_{E2}(t-e_i)$ are I-splines and their integrals all related to the second exposure. Further exposures can be incorporated in a similar way.

4. Penalized Log-Likelihood

The numbers of knots, which determine the numbers of M-spline basis functions that make up the age-specific and exposure-related relative incidence functions are chosen a priori. Maximizing the log-likelihood function (8) after choosing too large a number of knots over-fits the true curves, while selecting too small a number of knots leads to under-fitting overly smoothed curves. Therefore, to control the smoothness of the estimated functions we fix the numbers of knots at higher values than are believed to be enough to represent the functions and introduce roughness penalty terms to the log-likelihood function (8). We choose a roughness measure to be the sum of the square norms of the second derivatives of the age and exposure effect functions, following [10]. This leads to the penalized log-likelihood function

$$\begin{aligned} pl &= l(\boldsymbol{\alpha},\boldsymbol{\beta}) - \lambda_1\int\left\{\sum_{l=1}^{m_1}\alpha_l^2 M^{''}_{1l}(u)\right\}^2 du - \lambda_2\int\left\{\sum_{l=1}^{m_2}\beta_l^2 M^{''}_{2l}(u)\right\}^2 du \\ &= l(\boldsymbol{\alpha},\boldsymbol{\beta}) - \lambda_1\left\{(\boldsymbol{\alpha^2})^T\mathbf{A}_1\boldsymbol{\alpha^2}\right\} - \lambda_2\left\{(\boldsymbol{\beta^2})^T\mathbf{A}_2\boldsymbol{\beta^2}\right\} \end{aligned} \quad (9)$$

where $\boldsymbol{\alpha}$ is a vector of parameters $\alpha_1,\ldots,\alpha_{m_1}$, that define the age-specific relative incidence function and $\boldsymbol{\alpha^2} = \alpha_1^2,\ldots,\alpha_{m_1}^2$, $\boldsymbol{\beta}^2 = \beta_1^2,\ldots,\beta_{m_2}^2$ are parameters related to the exposure effect, $\mathbf{A}_1$ is an $m_1\times m_1$ matrix with (r,l) element $\int M^{"}_{1r}(u)M^{"}_{1l}(u)du$, $\mathbf{A}_2$ is an $m_2\times m_2$ matrix with (r,l) element $\int M^{"}_{2r}(u)M^{"}_{2l}(u)du$, $l(\boldsymbol{\alpha},\boldsymbol{\beta})$ is the log-likelihood function (8). λ_1 and λ_2 are non-negative smoothing parameters that control the trade-off between the model fit and the smoothness of the functions. The penalized log-likelihood function (9) is maximized, for fixed λ_1 and λ_2 values, to estimate the parameters related to age and exposure effects.

To estimate the parameters related to the age effect we need a constraint for $\psi(t)$ as it is a relative, not an absolute effect. In the standard SCCS model, $\psi(t)$ is set to 1 in a particular age group chosen as the reference. Here, a more readily interpretable constraint is $\int_a^b\psi(t)dt = 1$, which can be achieved by requiring $\sum_{l=1}^{m_1}\alpha_l^2 = 1$. The penalized log likelihood function (9) is maximized with constraint $\alpha_r = 1$, for $r = [(m_1+1)/2]$, then the parameters α_l are scaled to meet the constraint $\sum_{l=1}^{m_1}\alpha_l^2 = 1$.

4.1. Selection of Smoothing Parameters

We choose the smoothing parameters by maximizing approximate cross-validation scores [11]. λ_1 is first chosen by ignoring the exposure effect then λ_2 by ignoring the age effect. The smoothing parameter for the age effect λ_1 is chosen as described in [4] and the smoothing parameter for the exposure effect λ_2 is chosen as described in [5], see appendix A3 for details. After choosing the smoothing parameters the log-likelihood function (9) is maximized for fixed λ_1 and λ_2 values.

5. Calculation of confidence bands

Following Joly and Commenges [12] and O'Sullivan [11], approximate point wise 95% confidence bands for the exposure-related relative incidence function $\widehat{\omega}(\tau)$ may be obtained as follows. Let $\widehat{\boldsymbol{V}}$ denote the approximate covariance of $\widehat{\boldsymbol{\beta}}$, obtained from the negative of the inverted hessian of the penalized likelihood pl evaluated at the penalized maximum log likelihood estimates. Then

$$\widehat{\boldsymbol{W}} = 4\text{diag}(\widehat{\boldsymbol{\beta}})[\widehat{\boldsymbol{V}}]\left\{\text{diag}(\widehat{\boldsymbol{\beta}})\right\}^T$$

is the approximate covariance matrix of $\widehat{\boldsymbol{\beta}}^2$. Let $\boldsymbol{M_2(\tau)}^T = \{M_1(\tau), \ldots, M_{m_2}(\tau)\}$

Therefore the approximate confidence bands on $\widehat{\omega}(\tau) = \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(\tau)$ are

$$\widehat{\omega}(\tau) \pm 1.96\sqrt{\boldsymbol{M_2(\tau)}^T \widehat{\boldsymbol{W}}\, \boldsymbol{M_2(\tau)}} \tag{10}$$

The 95% coverage probabilities of these confidence bands in the context of the SCCS method were studied in a simulation study [5].

Alternatively, to ensure that the confidence bands lie above zero, they can be obtained on the log scale as

$$\widehat{\omega}(\tau)\exp\left\{\pm 1.96\sqrt{\boldsymbol{M_2(\tau)}^T \widehat{\boldsymbol{W}}\, \boldsymbol{M_2(\tau)}}\right\}$$

6. Simulation Study

To evaluate the performance of the new spline-based SCCS method and to compare it with the extensions made to the standard SCCS method in [4] and [5], we conducted a simulation study. Previous simulation studies presented in [4] and [5] showed that the use of splines has a better performance in terms of efficiency than the standard SCCS methods (when the incidence is not flat).

6.1. Design of the Simulation Study

The numbers of cases used in this simulation were $N = 100, 200, 1000$, each with ages at the start and end of the observation period of 0 and 730 days respectively. For each case, the risk period between the age at start of exposure c_i and age at end of exposure d_i was taken as 49 days. The baseline incidence was generated from a sine function, defined as $\lambda_0(t) \propto 8(\sin(0.01 \times t)) + 9$ at age t. The true age-related relative incidence function is presented in Panel a of Figure 1. Ages at start of exposure c_i, for $i : 1, \ldots, N$, were sampled within (0,730] from an exponential density with rate 0.003. The histogram of c_i is shown in Panel b of Figure 1.

Figure 1 about here

For the given age-related relative incidence function and distribution of age at exposure, we investigated four scenarios for the exposure-related relative incidence function, $\omega(t-c)$. These are illustrated in Figure 2: in Scenario 1 the exposure-related incidence increases and decreases beginning and ending at baseline level, in Scenario 2 the relative incidence is constant over the risk period, in Scenario 3 and 4 the relative incidence starts high and tails off to baseline, this is more gradual in Scenario 3. These functions take the value one outside the risk period $[c_i, d_i]$, that is when time since start of exposure $t-c<0$ or $t-c>49$. Without loss of generality we consider each case to have experienced only one event. The daily incidence rates within the observation period are evaluated as the product of the age-related relative incidence and the exposure-related relative incidence. An event day for each individual was generated from a multinomial distribution. The probability of an event at a given day within the observation period was computed as the incidence rate for that day divided by the sum of the rates for all the days within the observation period. For each scenario 100 data sets were simulated.

The data sets generated were analyzed in three ways:

1. SCCS with smooth age effect and parametric exposure effect (step function with 7 steps of length 7 days each) [4],
2. SCCS with parametric age effect (step function with 6 and 3 steps) and spline-based exposure effect [5], and
3. the spline-based SCCS, the new method proposed in this paper.

For the first method, seven exposure groups of length seven days between 0 and 49 were chosen to represent the exposure effect by a step function. For methods (1) and (3), to represent the age effect with a spline function 12 interior knots between the minimum of ages at the start of observation (zero) and the maximum of the ages at the end of observation periods (730) were chosen. The smoothing parameters for all the samples were chosen using the cross validation method.

For the second method, where age is represented with a piecewise constant function, six age groups with cut points at 0, 120, 240, 360, 480, 600 and 730 days were pre-specified. To represent the exposure effect with a spline function in methods (2) and (3), a nominal risk period of 49 days was chosen. 12 interior knots between zero and 49 were selected. The smoothing parameter of the exposure was chosen by the cross validation method for all the samples in the two methods. In addition, we fitted method (2), but with only three age groups with cut points at 0, 240, 480 and 730 days, to see how a change in age groups affects the results.

To compare the performance of the three methods in terms of estimating the age-specific relative incidence and the exposure related relative incidence we used the median of the integrated squared errors (MISE) and the standard deviation of the integrated squared errors (SD). For the age effect we constrained the relative incidence functions to be one at the maximum age to make the three methods comparable. The integrated squared error (ISE) for each sample is defined as

$$\left\{\psi_c(t)-\widehat{\psi}_c(t)\right\}^2,$$

where $\psi_c(t)$ is the true age-specific relative incidence constrained to be one at age 730 days and $\widehat{\psi}_c(t)$ is the estimated constrained relative incidence. After fitting the models for each sample we estimated the relative incidence at each day of age from 0 - 730 and approximated the ISE values as:

$$\sum_{t=0}^{730}\left\{\psi_c(t)-\widehat{\psi}_c(t)\right\}^2.$$

We then evaluated the MISE values as the median of the ISE values of the 100 samples in each scenario and the SD as the standard deviation of the ISE values. Similarly the ISE value for the exposure-related relative incidence is defined as

$$\int_c^d \left\{\omega(t-c)-\widehat{\omega}(t-c)\right\}^2 dt,$$

where $\omega(t-c)$ is the true exposure-related relative incidence function, and $\widehat{\omega}(t-c)$ is the estimated relative incidence function. The true exposure-related relative incidence functions used in the simulations are presented in Figure 2.

Figure 2 about here

6.2. *Results of the Simulation Study*

Table 1 presents the MISE and SD for estimating the age and exposure effects using the three methods. The method outlined in [5] was fitted twice for each generated data set using 6 and 3 age groups.

Table 1 about here

The results in Table 1 suggest that the new method performs well. As expected, the performance of all three methods in estimating the age and exposure related relative incidence curves reduces as the number of cases in the analysis decreases.

In estimating the age-specific relative incidence (RI) function the spline-based method has equivalent performance to method (1) with smooth age effect, and has better performance as compared to method (2) with six age groups when $N = 200, 1000$ (when $N = 100$ performance is equivalent). Performance in estimating the age RI is consistently poor for method (2) with three age groups.

In estimating the exposure-related relative incidence function, the fully spline-based method showed the highest performance as compared to both methods (1) and (2) when the sample size was large $N = 1000$. With sample sizes $N = 100, 200$, performance of the new method (3) and method (2) in estimating the exposure RI are somewhat variable, but broadly similar. With these smaller sample sizes MISE values are influenced by a handful of simulations where the estimated curves do not follow the trend of the true functions well, and these particular fits were consistently poor for all methods. When $N = 100, 200$, method (3) performs better than method (2) in 49.5% of all simulation runs, so we conclude that either is a reasonable approach when sample sizes are small.

The estimated age-related and exposure-related relative incidence functions along with their true curves are presented in Figures 3, 4, for scenarios 1 and 2, respectively with $N = 1000$ (the model with three age groups is not presented). Similar figures for scenarios 3 and 4, and for all scenarios with $N = 200$ are given in web appendix A. The curves related to the age effect are plotted by constraining the cumulative relative incidence at the maximum of the ages at the end of observation period to be one. The figures suggest that the spline-based method performs well in estimating both the age and exposure-related relative incidence curves. In all cases when $N = 1000$ the true functions are within the range of the estimated curves and the estimated curves equally follow the trend of the true functions. However there are some estimated exposure-related curves that over-fitted the true curve for scenario 2 when $N = 1000$, (Figure 4), where the true function is a constant. The optimal model for Scenario 2 would be method 1 with the number of exposure groups reduced to one.

Figure 3 about here

Figure 4 about here

7. Application

We illustrate the spline-based self-controlled case series method by applying it to data on measles, mumps and rubella (MMR) vaccine and febrile convulsions. The data set includes $2,389$ cases aged between 29 and 730 days with $3,826$ events. The data were collected in England and Wales in the period 1991-1994. To fit SCCS models the data should be listed one line per event and include the following variables: case identifier, age at start of the observation period, age at the end of observation period, age at start of exposure risk period, age at end of exposure risk period and age at event of interest.

We analyzed the data using the spline-based SCCS method developed in this paper where linear combinations of cubic M-splines are used to represent the age and exposure effects. The data had been analyzed previously using the standard SCCS method where age and exposure effects are piecewise constant functions [1]. This analysis used quarterly age groups

and two exposure risk periods, namely $6-11$ and $15-35$ days since the day of vaccination, yielding relative incidence (RI) estimates of 2.11 (1.21, 3.69) and 0.58 (0.33, 1.03) respectively. For comparison purposes we carried out three further standard SCCS analyses with varying exposure risk periods that are given in web appendix B. We also analyzed the data using the methods developed in [4] and [5]. We used seven exposure groups for the method where only age effect is represented by spline functions. In the method where only exposure is modelled using splines we used 22 age groups to represent age related relative incidence using piecewise constant function. Results of these analyses are presented in the Web Appendix B.

In fitting the spline-based SCCS model, for the MMR vaccine-related relative incidence function we chose a nominal risk period of 50 days. We used 12 equally spaced interior knots between 0 and 50. The smoothing parameter λ_2 for the exposure effect was chosen by the cross validation method and was found to be 3.725. For the age-related relative incidence, we used 12 interior knots between 29 and 730 and chose the smoothing parameter using the cross validation method. The value selected was 26176426. Then for the given values of the smoothing parameters, we maximized the spline-based SCCS penalized log-likelihood function (9). The estimated age and exposure-related relative incidence curves are presented in Figure 5.

Figure 5 about here

The dotted lines in Panel (b) of Figure 5 are approximate 95% confidence bands obtained by using a Bayesian-like technique [11] (10).

Panel (a) of Figure 5 shows the estimated age-related relative incidence function, where the cumulative age effect is constrained to have the value one at the maximum end of observation period. Panel (b) of the figure shows the relative incidence curve post MMR vaccine. From the figure, it can be seen that there is a significant increase in the risk of febrile convulsions from six to twelve days after exposure to MMR vaccine. Nine days after the vaccination the relative incidence of febrile convulsions peaks at around 2.5, which means the incidence of febrile convulsions is roughly 2.5 times higher on day nine after MMR vaccination than on any other day during the baseline periods before and after vaccination, after age effects have been taken into account. Day five and day thirteen after vaccination appear to be associated with a borderline significant inflated risk of febrile convulsions, and there is no significant increased risk in the periods 0-4 and 14-50 days after vaccination.

The basis for the choice of the first risk period in the original analysis [1] was that MMR vaccine contains attenuated live measles virus, and it was believed that it would take about a week for the measles virus to replicate to such levels as to cause symptoms and precipitate a convulsion. The second risk period, associated with Urabe strains of the mumps virus, was chosen based on other studies. It is of interest subsequently to look more broadly at the shape of the risk function in Figure 5 and thus see post hoc whether such a priori choices were justified, which appears to be the case, especially for the first risk period. It appears, however, that results of the original analyses were a little attenuated toward 0. Figure 5 helps us to understand why that might be so; quarterly age groups were inadequate to control of the strong age effect and perhaps the exposure risk periods failed to fully capture all periods with some excess exposure risk. Analysis (1) in appendix A5 captures age and exposure effects using piecewise constant functions better than the original analysis, but inefficiently has 10 risk periods and 21 age groups.

8. Final remarks

The method developed here combines the extensions to the standard SCCS method that modelled the age effect only using splines [4] and that modelled the exposure effect only using splines [5]. In this paper, the effects of both age and exposure in the SCCS model are represented by linear combinations of M-spline basis functions simultaneously. In other versions of the SCCS method either the age effect or the exposure risk period, or both, are represented by step functions that can yield biased estimates if the priori selected groups are poorly chosen or mis-specified. Our motivation was to develop a

new method that can avoid this.

It has been demonstrated that the use of splines is superior to control for strong age effects (or alternatively strong season effects) [4]. Standard methods involve step functions. However, to describe the risk function with sufficient flexibility, it is usually necessary to use many steps, which generally reduces efficiency as seen from the simulation studies. Our method seeks to address this problem in a rather natural manner via splines, which provide efficient estimates without sacrificing flexibility. We acknowledge that splines may over-fit age effects when they are truly flat or almost flat, but epidemiologists will generally have an informed idea of the strength of age effects a priori.

The new method also provides a useful description of the relative risk functions, without any prior assumptions about their shapes. This can then lead to more focused analyses to test hypotheses about potentially increased risks in specific time intervals, if required, in other data sets. Alternatively, our methods may be used to give a broader understanding of the shape/timing of risk after exposure and visualise how long it takes for risk returns to baseline levels, once a specific problem has been identified with, say, a particular vaccine.

Care must be taken to ensure that periods of inflated (or decreased risk) are not included in the baseline period. It is important that the risk period is not chosen to be too short, and those using this method for epidemiological research are advised that it is preferable that the risk period is either chosen to be a little too long, or that a washout period represented by a step function is included. Washout and pre-exposure 'risk' periods represented by step functions can easily be incorporated. It is not unusual for researchers to be unsure about how best to choose exposure risk and age periods when using the standard SCCS method. To explore this, or to achieve greater accuracy, a sensible approach is to define several contiguous risk periods or to increase the number of age groups. The fully-spline based method offers greater efficiency in comparison when fewer parameters need to be estimated overall. In addition it offers the advantage of representing the exposure risk with biologically plausible shapes, graphical displays of which may be of particular interest, and may help researchers explore the optimal choice of risk period visually.

Note that we do not envisage that spline-based exposure effects should replace step functions where the goal of a study is to quantify the excess risk (should any exist) for a clear pre-specified window of time after the start of an exposure, as splines do not produce an easily interpretable numerical summary in the same way. Rather, we envisage that spline-based exposure effects will be useful for exploratory or secondary analyses, either for hypothesis-generating studies, or as a check that exposure risk windows have been appropriately identified.

Development of the fully-spline based SCCS method was not trivial because the denominator of the log-likelihood function of the new method includes the integral of a product of two spline functions, namely the age-related and the exposure-related relative incidence functions. Rather than using numerical integration techniques, we evaluated this integral analytically using integration by parts. This required evaluation of the first, second and third integrals of an I-spline function, based on the definition of the integral of an M-spline [6].

Saarela [13], similar to [4], used cubic splines to take the age effect into account in case-base sampling methods [14]. In the self-matched case-base sampling, instead of using the whole observation period of the case, times in the observation period are randomly sampled and used in the denominator in order to completely avoid the integral in the denominator with minimal loss of information due to sampling. Piecewise constant estimates were used for the exposure effect [13]. Jensen [15] also used Splines in the case-time-control design to model changes in exposure probability. Our method uses splines to represent both age and exposure effects and uses the whole observation period of cases without any loss of information.

The simulation studies showed that the new method performs well compared with SCCS models with a single spline function with large samples, while performance was similar with smaller samples. Computation is a little slower than for the standard SCCS method, but analyses in this paper each took less than 10 minutes to fit. An R-package that includes the fully-spline based model is available from our SCCS website http://statistics.open.ac.uk/sccs/r.htm. The current version of the R package includes only the most basic new model with one risk period (though we plan to incorporate washout and pre-risk periods).

Possible future directions for this work have been considered. The tuning parameters of the age and exposure related relative incidence spline functions were chosen using a cross validation method, one possibility would be to consider choosing these parameters via a mixed effects representation. Another possibility would be to extend this method to more than one exposure (we have shown how multiple exposures can be included using splines in the current framework as long as there is no overlap in the risk periods). The current setting could allow additional exposures or time-varying covariates to be incorporated as piecewise constant functions. Similarly, two time-varying covariates could be modeled using splines while age is represented by a step function.

Acknowledgement

This research was supported by an MRC methodology grant MR/L009005/1, and PF was also supported by a Royal Society Wolfson research merit award.

References

1. Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* 1995; **51**: 228–235. DOI: 10.2307/2533328.
2. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine* 2006; **25**: 1768–1797. DOI: 10.1002/sim.2302.
3. Weldeselassie YG, Whitaker HJ, Farrington CP. Use of the self-controlled case series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection* 2011; **139**: 1805–1817. DOI: 10.1017/S0950268811001531.
4. Ghebremichael-Weldeselassie Y., Whitaker HJ, Farrington CP. Self controlled case series method with smooth age effect. *Statistics in Medicine* 2014; **33(4)**: 639–649. DOI: 10.1002/sim.5949.
5. Ghebremichael-Weldeselassie Y, Whitaker HJ, Farrington CP. Flexible modelling of vaccine effect in self-controlled case series models. *Biometrical Journal* 2016; **58(3)**: 607–622. DOI: 10.1002/bimj.201400257.
6. Ramsay JO. Monotone Regression Splines in Action. *Statistical Science* 1988; **3**: 425–461. DOI: 10.1214/ss/1177012761.
7. Farrington CP, Whitaker HJ. Semiparametric analysis of case series data. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 2006; **55**: 553–580. DOI: 10.1111/j.1467-9876.2006.00554.x.
8. Farrington CP, Whitaker HJ, Hocine MN. Case series analysis for censored, perturbed or curtailed post-event exposures. *Biostatistics* 2009; **10**: 3–16. DOI: 10.1093/biostatistics/kxn013.
9. Farrington CP, Anaya-Izquierdo K, Whitaker HJ, Hocine MN, Douglas I, Smeeth L. Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association* 2011; **106**: 417–426. DOI: 10.1198/jasa.2011.ap10108.
10. Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; **3(3)**: 433–443. DOI: 10.1093/biostatistics/3.3.433.
11. O'Sullivan F. Fast computation of fully automated log-density and log-hazard estimators. *Siam Journal on Scientific and Statistical Computing* 1988; **9**: 363–379. DOI: 10.1137/0909024.
12. Joly P, Commenges D. A Penalized Likelihood Approach for a Progressive Three-State Model with Censored and Truncated Data: Application to AIDS. *Biometrics* 1999; **55(3)**: 887–90. DOI: 10.1111/j.0006-341X.1999.00887.x.
13. Saarela O. A case-base sampling method for estimating recurrent event intensities. *Lifetime Data Analysis* 2016; **22**: 589-605. DOI: 10.1007/s10985-015-9352-x.
14. Saarela O, Hanley JA. Case-Base Methods for Studying Vaccination Safety. *Biometrics* 2015; **71(1)**: 42–52. DOI: 10.1111/biom.12222.
15. Jensen AKJ, Gerds TA, Weeke P, Torp-Pedersen C, Andersen PK. On the validity of the case-time-control design for autocorrelated exposure histories. *Epidemiology* 2014; **25(1)**: 110–113. DOI: 10.1097/EDE.0000000000000001.
16. de Boor C. 1978. *A Practical Guide to Splines*. New York: Springer-Verlag. 1995

A. Appendix

A.1. Derivatives of M-splines

M-splines are pieces of polynomials connected at points known as knots. M-splines of order q are defined as divided differences of truncated power functions [16], that is, for a given knot sequence $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$

$$M_l(t|q) = (-1)^q q[k_l, \ldots, k_{l+q}]T_t^q(k)$$

where $T_t^q(k)$ is a truncated power function of order q given by $T_t^q(k) = (t-k)_+^{q-1}$ and $[k_l, \ldots, k_{l+q}]T_t^q(k)$ is the q^{th} order divided difference of the function $T_t^q(k)$ at $k_l, \ldots, k_{l+q}$. Therefore, the first derivative of an M-spline function is

$$\frac{dM_l(t|q)}{dt} = (-1)^q q[k_l, \ldots, k_{l+q}]\frac{dT_t^q(k)}{dt}$$

and the derivative of a truncated power function of order q is given by

$$\frac{dT_t^q(k)}{dt} = (q-1)(t-k)_+^{q-2} = (q-1)T_t^{(q-1)}(k)$$

then using the definition of divided differences [16], we have

$$\frac{dM_l(t|q)}{dt} = (-1)^q q(q-1)\left\{\frac{[k_{l+1}, \ldots, k_{l+q}]T_t^{(q-1)}(k) - [k_l, \ldots, k_{l+q-1}]T_t^{(q-1)}(k)}{k_{l+q} - k_l}\right\}$$

$$\begin{aligned}
&= \frac{q}{k_{l+q} - k_l}(-1)^{q-1}(q-1)\left\{[k_l, \ldots, k_{l+q-1}]T_t^{(q-1)}(k) - [k_{l+1}, \ldots, k_{l+q}]T_t^{(q-1)}(k)\right\} \\
&= \frac{q}{k_{l+q} - k_l}\left\{(-1)^{q-1}(q-1)[k_l, \ldots, k_{l+q-1}]T_t^{(q-1)}(k) - (-1)^{q-1}(q-1)[k_{l+1}, \ldots, k_{l+q}]T_t^{(q-1)}(k)\right\} \\
&= \frac{q}{k_{l+q} - k_l}\left[M_l\left\{t|(q-1)\right\} - M_{l+1}\left\{t|(q-1)\right\}\right].
\end{aligned}$$

In general, the j^{th} derivative of an M-spline function of order q, $M_l(t|q)$, is

$$\frac{d^j M_l(t|q)}{dt^j} = \frac{q}{k_{l+q} - k_l}\left[\frac{d^{j-1}M_l\left\{t|(q-1)\right\}}{dt^{j-1}} - \frac{d^{j-1}M_{l+1}\left\{t|(q-1)\right\}}{dt^{j-1}}\right],$$

so the j^{th} derivative of a function which is a linear combination of M-spline basis functions, $f(t) = \sum_{l=1}^m \alpha_l M_l(t|q)$, can be given as

$$\frac{d^j f(t)}{dt^j} = \sum_{l=1}^m \alpha_l \frac{d^j M_l(t|q)}{dt^j}.$$

A.2. Integrals of I-splines

The second integral of an I-spline, the integral of $I_l^1(t|q)$, can be obtained in a similar way to the first integral described in the paper and is defined as, $I_l^2(t|q) = \int_o^t I_l^1(u|q)du$:

$$I_l^2(t|q) = \begin{cases}
0, & l > h \\
\sum_{m=l}^h \frac{(k_{m+q+1} - k_m)}{q+1}\sum_{n=m}^h \frac{(k_{n+q+2} - k_n)}{q+2}\int_o^t M_n(u|q+2)du, & h-q+1 \le l \le h \\
\frac{t^2}{2} - tk_{l+q} + \frac{k_{l+q}^2}{2} & \\
+\sum_{m=l}^h \frac{(k_{m+q+1} - k_m)}{q+1}\sum_{n=m}^h \frac{(k_{n+q+2} - k_n)}{q+2}\int_o^{k_{l+q}} M_n(u|q+2)du, & l < h-q+1.
\end{cases}$$

but $\int_o^t M_n(u|q+2)du$ and $\int_o^{k_{l+q}} M_n(u|q+2)du$ are I-splines of order $q+2$, therefore

$$I_l^2(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} (k_{r+q+3}-k_r) \frac{M_r(t|q+3)}{q+3}, & h-q+1 \le l \le h \\ \frac{t^2}{2} - tk_{l+q} + \frac{k_{l+q}^2}{2} + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} (k_{r+q+3}-k_r) \frac{M_r(k_{l+q}|q+3)}{q+3}, & l < h-q+1. \end{cases}$$

And the third integral of an I-spline, $I_l^3(t|q) = \int_o^t I_l^2(u|q)du$, is given as

$$I_l^3(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} \frac{(k_{r+q+3}-k_r)}{q+3} \sum_{v=r}^{h} (k_{v+q+4}-k_v) \frac{M_v(t|q+4)}{q+4}, & h-q+1 \le l \le h \\ \frac{t^3}{6} - \frac{t^2 k_{l+q}}{2} + \frac{tk_{l+q}^2}{2} - \frac{k_{l+q}^3}{6} + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} \frac{(k_{r+q+3}-k_r)}{q+3} \sum_{v=r}^{h} (k_{v+q+4}-k_v) \frac{M_v(k_{l+q}|q+4)}{q+4}, & l < h-q+1. \end{cases}$$

A.3. Selection of smoothing parameters

Denote the cross-validation scores [11] by $V_1(\lambda_1)$ and $V_2(\lambda_2)$,

$$V_1(\lambda_1) = \sum_i^N l_i(\widehat{\boldsymbol{\alpha}}_{-i}) \tag{11}$$

$$V_2(\lambda_2) = \sum_i^N l_i(\widehat{\boldsymbol{\beta}}_{-i}) \tag{12}$$

where $\widehat{\boldsymbol{\alpha}}_{-i} = \widehat{\boldsymbol{\alpha}}_{-i}(\lambda_1)$ is the maximum penalized likelihood estimator of $\boldsymbol{\alpha}$ (with the exposure effect excluded from the model) when individual i is removed, and l_i is the log likelihood contribution of individual i. Following [11], $V_1(\lambda_1)$ may be approximated by $\overline{V}_1(\lambda_1)$,

$$\overline{V}_1(\lambda_1) = l(\widehat{\boldsymbol{\alpha}}) - \text{tr}\left\{(\widehat{H}_1 - 2\lambda_1 \mathbf{S}_1)^{-1} \widehat{H}_1\right\}, \tag{13}$$

where tr(X) is the trace of a matrix X, $l(\widehat{\boldsymbol{\alpha}})$ is the log-likelihood function in Equation (8) where no exposure effect is included and evaluated at the maximum penalized likelihood estimates $(\widehat{\boldsymbol{\alpha}})$. $\widehat{H}_1 = \frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}(\widehat{\boldsymbol{\alpha}})$ is the log-likelihood part of the Hessian of the penalized log-likelihood evaluated at the penalized maximum likelihood estimates $\widehat{\boldsymbol{\alpha}}$. The matrix $\mathbf{S}_1$ depends on the expression for $g(\alpha_l)$. If $g(\alpha_l) = \alpha_l$ then $\mathbf{S}_1 = \mathbf{A}_1$ [12], however here we take $g(\alpha_l) = \alpha_l^2$. Therefore, $\mathbf{S}_1 = 4\left\{\mathbf{A}_1 \circ (\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right\} + 2\left\{\text{diag}(\mathbf{A}_1 \boldsymbol{\alpha}^2)\right\}$ [4], where $\circ$ is the Hadamard product of matrices. Similarly, to choose the smoothing parameter of the exposure-related relative incidence function, $V_2(\lambda_2)$ can be approximated as

$$\overline{V}_2(\lambda_2) = l(\widehat{\boldsymbol{\beta}}) - \text{tr}(\left\{(\widehat{H}_2 - 2\lambda_2 \mathbf{S}_2)^{-1} \widehat{H}_2\right\}, \tag{14}$$

where $l(\widehat{\boldsymbol{\beta}})$ is the log-likelihood (8) taking no age effect into consideration, $\widehat{H}_2 = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}})$ is the Hessian when no age effect is included and $\mathbf{S}_2 = 4\left\{\mathbf{A}_2 \circ (\boldsymbol{\beta}\boldsymbol{\beta}^T)\right\} + 2\left\{\text{diag}(\mathbf{A}_2 \boldsymbol{\beta}^2)\right\}$.

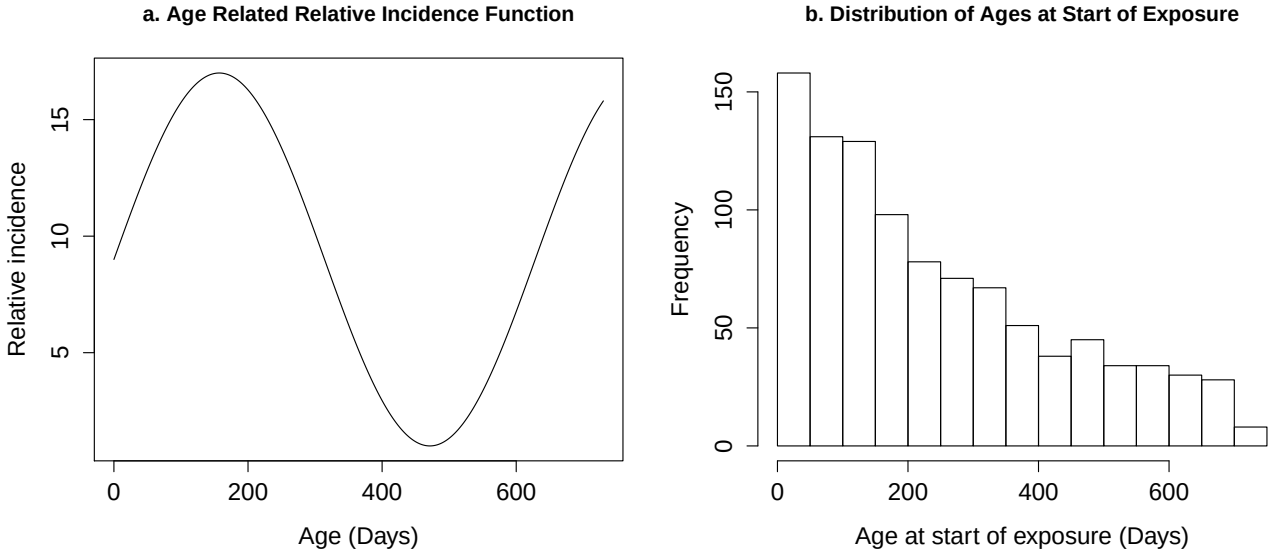

Figure 1. *True age-related relative incidence function in Panel* (a) *and distribution of ages at start of exposure in Panel* (b)*, which were used to simulate data sets.*

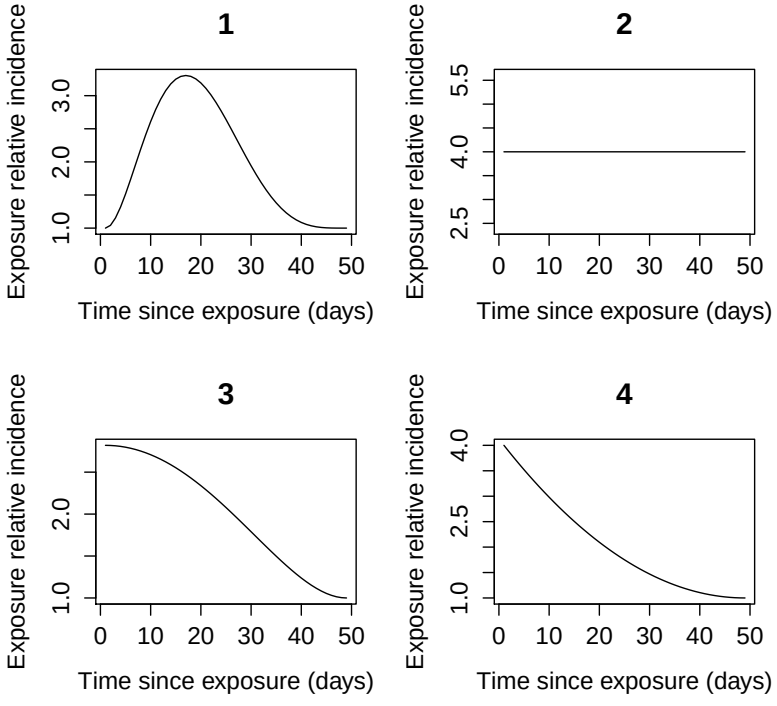

Figure 2. *True exposure-related relative incidence functions.*

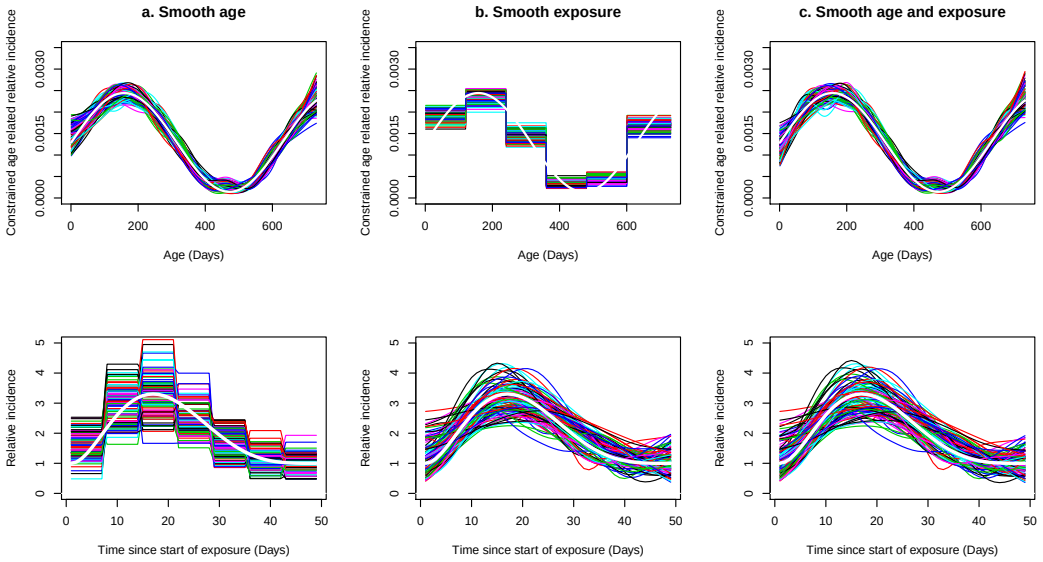

Figure 3. *Estimated relative incidence curves for scenario* 1*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

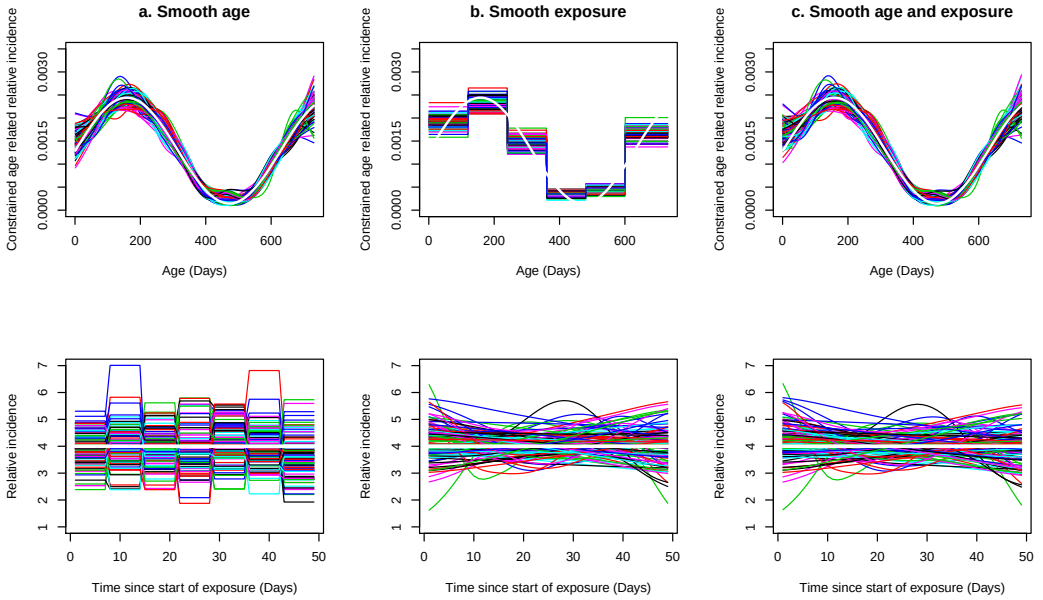

Figure 4. *Estimated relative incidence curves for scenario* 2*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

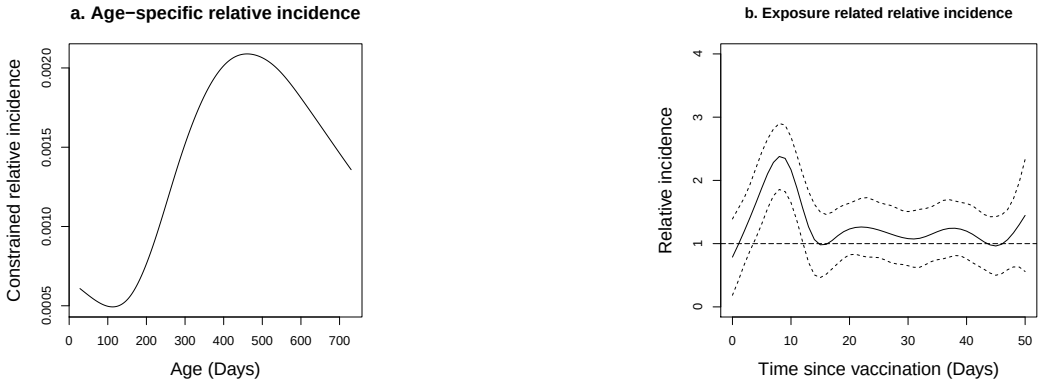

Figure 5. *Relative incidence curves estimated by fitting spline-based SCCS. Panel (a) shows the estimated constrained age-related relative incidence function Panel (b) represents estimated exposure-related relative incidence curve (solid line) along with 95% confidence bands denoted by the dashed lines.*

Table 1. *Median integrated squared errors (MISE) and standard deviation (SD) obtained from the three spline-based SCCS methods: (1) SCCS with smooth age effect, (2) SCCS with smooth exposure effect (twice with 6 and 3 age groups) and (3) SCCS with both age and exposure effects represented by splines. Each simulated data set was fitted by the three methods using a nominal risk period of* 49 *days. The true age-specific relative incidence function was generated from a sine function.*

		Method 1	Method 2 6 age groups	Method 2 3 age groups	Method 3
Scenario	Effects	MISE (SD)	MISE (SD)	MISE (SD)	MISE (SD)
		Sample size = 1000			
1	Exposure	11.914 (6.431)	6.808 (4.749)	6.848 (4.415)	6.571 (4.853)
	Age	0.011 (0.005)	0.069 (0.005)	0.243 (0.003)	0.011 (0.005)
2	Exposure	23.747 (13.516)	9.381 (11.408)	8.793 (9.440)	9.751 (10.792)
	Age	0.011 (0.006)	0.069 (0.006)	0.242 (0.003)	0.013 (0.006)
3	Exposure	8.686 (5.623)	3.411 (6.059)	3.721 (5.141)	3.079 (5.669)
	Age	0.010 (0.005)	0.069 (0.004)	0.243 (0.003)	0.011 (0.006)
4	Exposure	10.808 (4.506)	4.718 (3.720)	4.419 (3.902)	4.322 (3.562)
	Age	0.012 (0.006)	0.069 (0.004)	0.243 (0.003)	0.013 (0.007)
		Sample size = 200			
1	Exposure	45.885 (32.155)	33.044 (23.188)	31.545 (20.203)	32.689 (21.478)
	Age	0.035 (0.023)	0.084 (0.021)	0.246 (0.013)	0.037 (0.024)
2	Exposure	100.407 (88.300)	39.055 (65.633)	32.484 (58.444)	35.951 (63.845)
	Age	0.041 (0.023)	0.089 (0.026)	0.249 (0.013)	0.044 (0.024)
3	Exposure	45.167 (21.899)	14.031 (22.259)	13.131 (20.222)	14.270 (20.990)
	Age	0.040 (0.024)	0.090 (0.025)	0.248 (0.012)	(0.042 0.025)
4	Exposure	40.606 (30.048)	12.019 (22.297)	11.975 (20.671)	12.599 (21.280)
	Age	0.038 (0.023)	0.086 (0.023)	0.247 (0.015)	0.041 (0.023)
		Sample size = 100			
1	Exposure	92.438 (82.785)	50.336 (63.162)	50.299 (62.533)	48.205 (64.666)
	Age	0.078 (0.042)	0.117 (0.045)	0.257 (0.028)	0.080 (0.042)
2	Exposure	221.200 (176.226)	81.272 (92.604)	81.994 (96.788)	74.701 (88.620)
	Age	0.074 (0.036)	0.114 (0.049)	0.258 (0.026)	0.076 (0.037)
3	Exposure	86.171 (78.075)	25.076 (44.138)	26.028 (47.885)	24.744 (44.075)
	Age	0.073 (0.038)	0.114 (0.043)	0.256 (0.023)	0.075 (0.039)
4	Exposure	86.939 (62.777)	26.260 (47.530)	(26.671 48.879)	25.829 (44.892)
	Age	0.071 (0.044)	0.107 (0.044)	0.256 (0.030)	0.072 (0.045)